The Accelerating Growth of Online tagging Systems

Lingfei Wu

Dept. of Media and Communication, City University of Hong Kong, Hong Kong, China

Abstract. Research on the growth of online tagging systems not only is interesting in its own right, but also yields insights for website management and semantic web analysis. Traditional models that describing the growth of online systems can be divided between linear and nonlinear versions. Linear models, including the BA model (Brabasi and Albert, 1999), assume that the average activity of users is a constant independent of population. Hence the total activity is a linear function of population. On the contrary, nonlinear models suggest that the average activity is affected by the size of the population and the total activity is a nonlinear function of population. In the current study, supporting evidences for the nonlinear growth assumption are obtained from data on Internet users' tagging behavior. A power law relationship between the number of new tags (F) and the population (P), which can be expressed as $F \sim P^{\gamma}$ ($\gamma > 1$), is found. I call this pattern accelerating growth and find it relates the to time-invariant heterogeneity in individual activities. I also show how a greater heterogeneity leads to a faster growth.

1 Introduction

In the research literature, there are two kinds of models describing the growth of online social systems: linear and nonlinear. Linear models suggest that the expected value of individual activities M is a constant independent of active population P. As a consequence, the total amount of activity $F = MP$ is a linear function of P. For example, in the BA model for social network evolution [1], the average

degree is a constant parameter. On the contrary, nonlinear models argue that the size of the population has an effect on individual activities [2], which leads to a nonlinear relationship between F and P. The assumption of nonlinear growth is supported by findings in different online activities including game playing [3], resource recommendation [4], collaborative programming [5] and tagging [6]. Moreover, evidence on the nonlinear nature of human collective

behavior is also seen in many offline systems, such as cities and countries [7][8], academic groups [9], sexual networks [10], and so on.

In the current study, I investigate two online tagging systems and obtain supporting evidence for the nonlinear growth assumption. I find that within a tagging system, the number of new tags F is a power law function of the active population P with a constant exponent γ despite the daily fluctuation of tags and population. Since $\gamma > 1$ means that the number of new tags grows faster than the population does, the pattern can be named as "accelerating growth" [2][11][12]. It is observed that, the growth rate γ positively relates to the heterogeneity in individual tagging activities $1/\beta$, in which β is the exponent of power law distribution of individual activities (Eq.3). The heterogeneity $1/\beta$ is found to remain constant over time within a system, but differs across systems.

This paper is organized as follows. In section 2, I introduce the empirical accelerating growth patterns in two online tagging systems. In section 3, I show how the growth rate γ relates to the heterogeneity of individual activities $1/\beta$, and validate their relationship by empirical data and simulation. Finally, I briefly summarize the findings and discuss possible applications of accelerating growth in the informational industry.

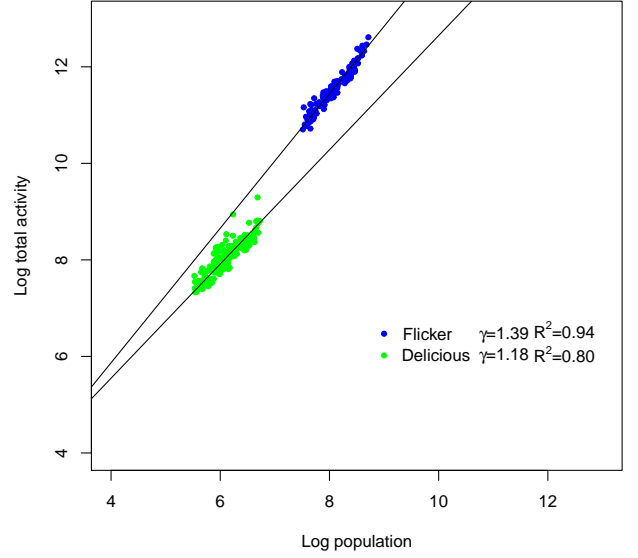## 2 Accelerating growth of online tagging systems

Fig. 1: Accelerating growth in online tagging systems. Different datasets are marked in points of different colors (blue for Flickr and green for Delicious).The x axis is the log of active population in a day and the y axis is the log of total activity in the day. The orthogonal log-log regression lines are also shown.

I investigate the tagging activities of 77,917 users on Flickr and 5,147 users on Delicious. The data from Flickr and Delicious are collected by systematically crawling the record of all active users [13]. In the two systems, I define P as the number of active users in a day, and F the number of new tags generated by the users. Please refer to Table 1 for the detailed information of the two systems.

The power law relationships between F and P in two systems, which can be expressed as

$$F \sim P^{\gamma} \tag{1}$$

Table 1: The empirical values of accelerating growth rate in online tagging systems..

Activity	System	γ	θ	95% CI of γ	Adjusted R^2	N of Days	Web Address
Book-tagging	Delicious	1.18	0.18	[1.11, 1.31]	0.80	171	delicious.com
Photo-tagging	Flickr	1.39	0.39	[1.32, 1.46]	0.94	120	flickr.com

or

$$Log(F) = \gamma Log(P) \qquad (2)$$

are found. Orthogonal regression [14] is used to estimate the value of γ in eq. (2). We use orthogonal regression but not ordinary least-squares regression because the latter tends to overstate the effect of outliers (data points with large variance). It is observed that the empirical values of growth rate γ in both systems are greater than 1. This finding of accelerating growth is consistent with results of previous studies [3][4][5][6]. According to eq. (1), it is easy to know that there is also a power law relationship between the average number of tags M posted by a user and the population, since $M = F/P \sim P^{\gamma-1} \sim P^{\theta}$. As $\gamma > 1$ and $\theta > 0$ in the data (Table 1), the average number of tags is not a constant; instead, it increases with population.

Comparing the θ in data with those reported in previous studies sheds lights on the accelerating nature of human collective behavior [7][8][9][15]. The phenomenon that average activities (e.g., average salary or average walking speed of pedestrians) increase with city population has been well studied [7][15], and is used to explain the different paces of urban and rural life. Similar patterns also include GDP per capita that increases with country size [8] and the average length of references in publications that increases with the scale of scientific collaboration network [9]. However, most of the θ observed offline are not greater than their counterparts online. For instance, the average activity usually scales to the system size with a θ range from 0 to 0.35 in offline systems [7][8] whereas θ found in my study can be greater than 0.38 (in Flickr). Given the population of a system, a larger θ means higher productivity in generating activities. Therefore we are reasonable to conjecture that online social systems, if are utilized appropriately, may be more "productive" than offline ones.

Then, why is there accelerating growth and what determines the value of θ (or γ)? Although several models have been proposed to answer the questions [16][17][18], there is still lack of a unified framework explaining the origin of accelerating growth. In the current study, I find that the accelerating growth relates to the stable (or time-invariant) heterogeneity in individual activities. In the next section, I will analytically show how heterogeneity gives rise to accelerating growth. Please note that this study does not aim to give an exclusive and unified framework towards the origins of all accelerating growth patterns, although I hope my model may contribute to such a framework in one way or another.

3 Heterogeneity and accelerating growth

Many models have been proposed to explain the nonlinear growth phenomenon [16][17][18][19], but few can be generalized across different fields of study. Some ecologists suggest that nonlinear growth originates from the self-similar network structure of biological individuals [17] or metabolic networks that minimize transportation cost [18]. Scholars in linguistics suggest that a skewed distribution of word occurrence leads to non-linear growth in word vocabulary [16]. Similar ideas are also seen in studies on the growth of family name [19] and open source software [5]. In the current research, I analytically attribute accelerating growth of online tags to the time-invariant heterogeneity of individual tagging activities in the system. The heterogeneity is quantified as the $1/\beta$, in which β is the exponent of power law distribution of individual activities. The positive correlation between heterogeneity $1/\beta$ and accelerating growth rate γ is verified by empirical data and numerical simulations.

3.1 Time-invariant power law distribution in online tagging systems

To find out why the average number of tags increases with population, I explore the daily distributions of user tagging activities. It turns out the distributions approach a straight line in log-log axis, suggesting a model of power law distribution. I use f to denote number of tags and $n(f)$ the number of users that generate so many tags. Then the power law probability distribution function (PDF) of f be expressed as

$$n(f) = C_t f^{-\beta} \tag{3}$$

In which C_t varies over time. In Eq.3 the value of β is set to be greater than one, since the cumulative distribution function (CDF) of power law distribution (Eq.3), called Pareto distribution [20], is also a power law distribution with an exponent $-\alpha = 1 - \beta < 0$.

In Figure 2, I draw 3 samples of daily distributions of individual activities on log-log axes for both systems. It is observed that, when population increases, the distributions move toward the right hand in parallel. Hence, I conjecture that C_t in eq. (3), which determines the intercept of the regression line, increase with system size, but β, which decides the slope of the line, is a size invariant constant. In other words, all the daily distributions should collapse to one theoretical curve if the variance of system size is controlled. I call the theoretical curve time-invariant power law distribution, since its slope does not change over time.

What makes the system maintain a time-invariant power law distribution in the temporal evolution? This is a question calling for further exploration. However, this phenomenon is non-trivial, since it means that the heterogeneity of the system does not change over time, which will give rise to accelerating growth. But before I discuss the connection between heterogeneity and accelerating growth, I would like to derive the function of time-invariant distribution at first.
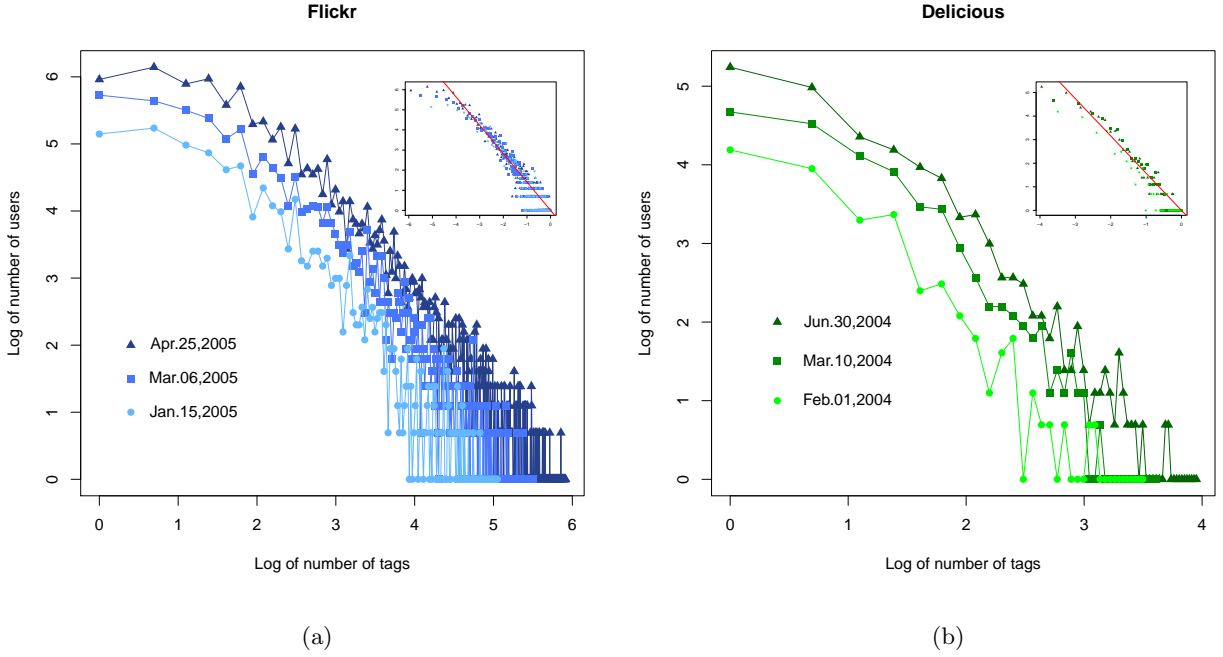
Fig. 2: Three examples of daily distributions of user tagging activities. The x axis is the log of number of tags and the y axis is the log of number of users. The rescaled distributions and the theoretical power law models with $\beta = 1.41$ (Flickr) and $\beta = 1.58$ (Delicious) are shown in the insets.

As in tagging systems, there is usually only one user who generates the maximum activity, that is

$$n(f_{max}) = 1 \approx C_t f_{max}^{-\beta} \rightarrow C_t = f_{max}^{\beta} \qquad (4)$$

With eq. (4), we can rewrite eq. (3) as

$$n(f) = (f/f_{max})^{-\beta} \qquad (5)$$

eq. (5) is the function of time-invariant power law distribution. As mentioned, all empirical daily distributions are supposed to collapse to the curve predicted by eq. (5) after rescaling. Therefore in both systems under study I rescale daily distributions together before estimating the values of β. According to the large values of adjusted R^2

Table 2: Parameters in time-invariant power law model of online tagging systems.

System	β	95% CI of β	Adjusted R^2	N of Days
Delicious	1.58	[1.57, 1.59]	0.90	171
Flickr	1.41	[1.40, 1.41]	0.90	120

shown in Table 2, the assumption about time-invariant distribution is plausible.

3.2 From power law distribution to accelerating growth

Given a distribution function, the relationship between the population $P = \int n(f)df$ and the total number of tags $F = \int n(f)f df$ is fully predictable. With the time-
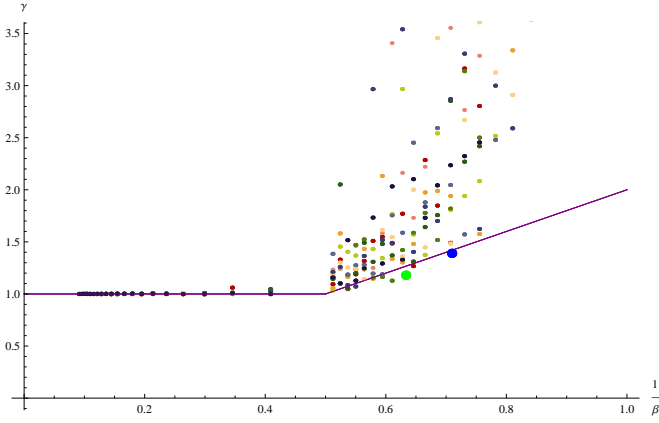
Fig. 3: The relationship between heterogeneity $(1/\beta)$ and growth rate (γ). The analytical function is draw in purple line, the results of numerical simulations are plotted in small dots whose color varies with C, and the empirical results of two tagging systems are marked in green (Delicious) and blue (Flickr) dots with large size. In the numeric simulation based on power law PDF $f(x) = C^{\beta-1}x^{-\beta}(\beta-1))$, C is an integer in $1 \leq C \leq 10$, and β is a fraction in $(1 < \beta \leq 10)$. For each value of C, 40 different values of β are selected. Therefore, there are 400 small dots in the figure.

invariant power law distribution shown above eq. (5), the time-invariant relationship between P and F can be derived as follows.

As $F_{max} \gg 1$ in empirical data,

When $\beta > 2$,

$$P = \int_1^{f_{max}} n(f)\,\mathrm{d}f = \frac{f_{max}}{1-\beta} - \frac{f_{max}^\beta}{1-\beta} \approx \frac{f_{max}^\beta}{\beta-1} \quad (6)$$

$$F = \int_1^{f_{max}} n(f)f\,\mathrm{d}f = \frac{f_{max}}{2-\beta} - \frac{f_{max}^\beta}{2-\beta} \approx \frac{f_{max}^\beta}{\beta-2} \quad (7)$$

Therefore $F \sim P$, $\gamma \approx 1$, $\theta \approx 0$. The greater β is, the more closely γ approaches to 1.

When $\beta = 2$,

$$P = \int_1^{f_{max}} n(f) \approx f_{max}^2 \approx \int_1^{f_{max}} n(f)f = F \quad (8)$$

Therefore $F \sim P$, $\gamma \approx 1$, $\theta \approx 0$.

Similarly, when $1 < \beta < 2$, we can derive that

$$P \approx \frac{f_{max}^\beta}{\beta-1} \quad (9)$$

$$F \approx \frac{f_{max}^2}{2-\beta} \quad (10)$$

Therefore $F \sim P^{\frac{2}{\beta}}$, $\gamma \approx \frac{2}{\beta}$, $\theta \approx \frac{2}{\beta}-1$. And smaller β leads to greater γ.

To sum up, we get

$$\gamma = \begin{cases} 2/\beta & \text{if } 1 < \beta < 2 \\ 1 & \text{if } \beta \geq 2 \end{cases} \quad (11)$$

According to eq. (11), if a system maintains a time-invariant power law distribution with an exponent $\beta < 2$, it will show property of accelerating growth. As power law distribution is famous for its high heterogeneity (i.e., quantitative difference in individual activities in the system), and smaller β means higher heterogeneity, we can regard $1/\beta$ as an index for heterogeneity and claim that greater heterogeneity gives rise to a faster growth (greater γ). As shown in Fig. 3 and Table 3, the theoretical function of γ on $1/\beta$ (eq. (11)) is consistent with empirical data and simulation results.

4 Discussions

In this paper, I discuss the accelerating growth of two online tagging systems, noting that it is the time-invariant heterogeneity of individual tagging activities that gives

Table 3: Comparison between empirical (γ) and theoretical (γ') accelerating growth rate

System	β	$\gamma' = 2/\beta$	95% CI of γ
Delicious	1.58	1.27	[1.11, 1.31]
Flickr	1.41	1.42	[1.32, 1.46]

rise to such a accelerating growth. Further, I analytically derive the function of accelerating growth rate γ on heterogeneity $1/\beta$, which is verified by empirical data and numerical simulations.

The findings of this paper support non-linear models of collective human behavior, which has been widely acknowledged by previous studies [2][3][4][5][6]. Moreover, the relationship between heterogeneity and accelerating growth may not only exist in online tagging systems, but also can be found in other online social systems, or even offline social systems.

Accelerating growth, once proved to be widely exist in online social systems, can find its applications in website management or semantic web analysis. For instance, by analyzing web traffic data, we can predict the traffic on a website with a given active population. Hence this pattern will help website masters plan their web server capacity accordingly. Also, we can compare γ between websites with equivalent functions and benchmark the most efficient one for other websites.

Some questions left points out the direction for the future studies. For example, why can online tagging systems maintain a time-interval power law distribution? Why C_t in eq. (3) is only determined by system size? These are all interesting questions worth further exploration.

Acknowledgements

References

1. A.L. Barabasi, R. Albert, Science **286**, 509 (1999)

2. J. S. Mattick, M. J. Gagen, Science **307**, 856 (2005)

3. T. Henderson, S. Bhatti, *Proc. the 9th ACM International Conference on Multimedia*(Ottawa,2001), Vol.1, p. 212

4. F. Wu, D. M. Wilkinson, B. A. Huberman, *Proc. the 11th IEEE International Conference on Computational Science and Engineering*(Vancouver, 2009), Vol.4, p. 409

5. C. J. Tessone, M. M. Geipel, F. Schweitzer, arXiv:1007.1330 (2010).

6. C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, V. Loreto, PNAS **106**, 10511 (2009)

7. L. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, G. B. West, PNAS **104**, 7301 (2007)

8. J. Zhang, T. Yu, Physica A **389**, 4887 (2010)

9. P. L. Krapivsky, S. Redner, Phys. Rev. E **71**, 36118 (2005)

10. B. F. de Blasio, A. Svensson, F. Liljeros, PNAS **104**, 10762 (2007)

11. M. A. Serrano, M. Boguna, A. Diaz-Guilera, Phys. Rev. Lett. **94**, 038701 (2005)

12. D. M. D. Smith, J.-P. Onnela, N. F. Johnson, New J. Phys. **9**, 181 (2007)

13. O. Gorlitz, S. Sizov, S. Staab, *Proc. the 7th international conference on Peer-to-peer systems*(Tampa Bay,2008), Vol.1, p. 19

14. M. Rodriguez, E. Langre, B. Moulia, Am. J. Bot. **95**, 1523 (2008)

15. S. Nordbeck, Geogr. Ann. B **53**, 54 (1971)

16. L. Lu, Z. K. Zhang, T. Zhou, PLoS ONE 5(12): e14139, doi:10.1371/journal.pone.0014139 (2010)

17. G. B. West, J. H. Brown, B. J. Enquist, Science **276**, 122 (1997)

18. J. R. Banavar, A. Maritan, A. Rinaldo, Nature **399**, 130 (1999)

19. S. Miyazima, Y. Lee, T. Nagamine, H. Miyajima, Physica A **278**, 282 (2000).

20. M. E. J. Newman, Contemp. Phys. **46**, 323 (2005)